不确定 NNSB-OPTICS 聚类算法在滑坡危险性预测中的研究与应用 *

毛伊敏，陈华彬，李忠利，张灿龙

(江西理工大学 信息工程学院, 江西 赣州 341000)

摘　要：针对滑坡危险性预测中降雨等不确定因素不能有效刻画及处理和现有的 OPTICS-PLUS 聚类算法需要设置密度阈值、时间复杂度高等问题进行了研究，为了提高滑坡危险性预测准确率，提出一种不确定 NNSB-OPTICS 聚类算法并应用于滑坡预测中。首先对 OPTICS-PLUS 算法扩张策略进行优化，避免了人工设置密度阈值，提高了算法效率；然后根据降雨量数据的分布特征，综合 EW 型距离公式和云模型理论，提出 EC 型距离公式，有效处理不确定数据降雨量；最后将不确定 NNSB-OPTICS 聚类算法应用于延安市宝塔区滑坡危险性预测中，建立滑坡危险性预测模型，滑坡预测精度达到 89.7%。实验结果表明，该方法能够有效提高滑坡危险性预测精度，具有较高可行性。

关键词：滑坡；危险性预测；不确定数据；OPTICS 算法

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.06.0653

Research and application of uncertain NNSB-OPTICS clustering algorithm in landslide hazard prediction

Mao Yinmin, Chen Huabin, Li Zhongli, Zhang Chanlong

(*School of Information Engineering Jiangxi University of Science & Technology*, *Ganzhou Jiangxi* 341000, *China*)

Abstract: Since the rainfall and other uncertainties are difficult to obtain and effectively deal with in landslide hazard prediction, and the existence of setting density threshold and high time complexity in the OPTICS-PLUS algorithms, in order to improve the prediction accuracy, this paper proposed an uncertainty NNSB-OPTICS clustering algorithm and applied to landslide prediction. Firstly, the expansion strategy of OPTICS-PLUS algorithm is optimized, which avoids the manual setting of density threshold and improves the efficiency of the algorithm. Then, according to the distribution characteristics of rainfall data, combined with EW distance formula and cloud model theory, this paper puts forward EC distance formula, can deal with the uncertain rainfall data effectively. Finally, the uncertain NNSB-OPTICS clustering algorithm is applied to predict landslide hazard in Baota district of Yan'an city and the landslide prediction accuracy reaches into 87.9%. The experimental results show that this method can effectively improve the accuracy of landslide prediction and has high feasibility.

Key Words: landslide; hazard prediction; uncertain data; OPTICS algorithm

0　引言

滑坡是分布最为广泛、发生最为频繁的地质灾害之一，给人类生存和发展带来了严重的威胁。滑坡的形成受多种因素的影响，不仅包括地形地貌、地层岩性和坡体结构等基本因素，还包含具有很大不确定性的降雨和人类活动等诱发因素[1]。在这些因素相互综合作用下，滑坡发生极具复杂性和不确定性，给滑坡危险性预测带来很大困难。

聚类分析是数据挖掘中的关键技术，能够在无先验样本的情况综合考虑影响滑坡发育的众多关键因素，并通过这些因素之间的相似性对滑坡数据对象进行分类，提取出潜在的有用信息[2]，因而越来越多学者借助聚类算法在滑坡预测中展开研究。文献[3]运用模糊 K 均值算法对道路滑坡和非道路滑坡的地形地貌进行分类，结合 GIS 技术建立了美国爱达荷州清水国家森林公园的滑坡危险性概率图，证明该方法能够准确预测道路相关滑坡，对道路规划有重要指导意义。文献[4]采用 K-means 聚类法对汶川灾区泥石流滑坡易发性进行划分，获得五个子类，根据专家经验对五个子类危险等级进行判定，研究表明 K-

基金项目：国家自然科学基金资助项目（41530640，41362015，41562019）；江西省自然科学基金资助项目（20161BAB203093）；江西省教育厅科技项目（GJJ151531）

作者简介：毛伊敏（1970-），女，新疆伊犁人，教授，主要研究方向为数据挖掘、地理信息系统（mymlyc1@163.com）；陈华彬（1993-），男，硕士研究生，主要研究方向为数据挖掘；李忠利（1991-），男，硕士研究生，主要研究方向为数据挖掘；张灿龙（1990-），男，硕士研究生，主要研究方向为数据挖掘.

means 聚类算法划分危险性等级的预测结果与实际情况一致性较高，划分效果较好。文献[5]以湖北省巴东县滑坡灾害调查资料为基础，选择具有代表性的滑坡灾害影响因素作为危险性评价指标，采用熵权法和 APH 法获取各评价因子权重值应用到 K-means 聚类算法中对 86216 个预测单元的滑坡危险等级进行预测，预测结果与当地滑坡灾害实际情况基本吻合，能够对多属性数据进行处理，具有一定的实用价值。文献[6]以三峡库区万州区为研究对象，选取对滑坡影响较大的 7 个致灾因子作为评价指标，使用滑坡面积比与分级面积比曲线对指标因子分级，然后使用 K-means 聚类法对易发性结果进行分级，并基于 GIS 平台建立易发性区划图，获得令人满意的预测精度。文献[7]分析山体滑坡的空间分布和变形因素提取潜在中心，使用潜在中心描述的云模型理论对 K-means 聚类算法进行改进，根据数据点隶属度对数据对象进行聚类分析，并将改进后的算法应用于三峡库区滑坡预测中，证明该方法能够对区域滑坡危险性进行很好的预测。文献[8]采用两阶段分析提取主要属性和阈值，计算对滑坡发生影响的熵值，使用粒子群思想优化 K-means 算法，解决了 K-means 算法容易陷入局部最优的问题，以台湾苗栗山池国家公园为研究区绘制敏感度图，实验证明了改进的 K-means 算法具有更高的预测精度。这些聚类算法在滑坡预测中取得了一定的成果，但远未达到让人满意的程度，存在两个方面原因：a)降雨不确定因素是滑坡发生的重要因素之一，这些聚类算法侧重于对连续数值属性和离散型数值属性的处理，对滑坡预测中的不确定数据降雨量不能进行有效刻画；b)以上传统聚类算法需要预先确定聚簇数目 k 和聚类中心，对于分布不均匀的数据无法有效处理，因此对滑坡数据集的聚类效果并不理想。由于存在以上原因，传统聚类算法的滑坡危险性预测精度不够高，因此需要探索一种新的方法，在适用于分布不均匀数据聚类的同时能够有效处理不确定数据降雨量，进一步提高滑坡危险性预测精度。

OPTICS-PLUS 聚类算法[9]是一种基于密度的聚类算法，相比于基于划分的 K-means 算法更加适用于滑坡预测中分布不均匀的数据聚类，且在聚类过程中采用了一次聚类结果重组织策略，使生成的可达图结构更加清晰，聚类准确率较高。但 OPTICS-PLUS 算法仍存在难以有效刻画降雨量，需要用户输入密度阈值，难以避免人的主观性和随意性对滑坡预测结果的影响，时间复杂度较高，不适合对大规模滑坡数据集聚类等不足。对此，本文在 OPTICS-PLUS 算法基础上进行改进，对算法的扩张策略进行优化，提出一种基于最近邻搜索的 OPTICS 算法（nearest neighbor search based OPTICS, NNSB- OPTICS）；然后根据降雨量数据的分布特征，将 EW 型距离公式[10]和云模型理论[11]相结合提出 EC 型距离公式，将 EC 型距离公式引入 NNSB-OPTICS 算法中，提出不确定 NNSB-OPTICS 算法，解决了滑坡预测中不确定数据降雨量难以有效刻画的问题。最后将不确定 NNSB-OPTICS 算法应用于延安市宝塔区滑坡危险性预测中，建立滑坡危险性预测模型，证明了算法可行性和有效性。

1　不确定 NNSB-OPTICS 聚类算法

1.1　NNSB-OPTICS 聚类算法设计

OPTICS-PLUS 算法是对 OPTICS 算法[12]的一种改进，解决了 OPTICS 算法因贪心搜索策略导致稀疏点不能有效聚类的问题，聚类准确率较高。但 OPTICS-PLUS 算法仍需要用户输入密度阈值，难以避免人的主观性和随意性对滑坡预测结果的影响，且时间复杂度较高，不适合对大规模滑坡数据集聚类。

为此本文在 OPTICS-PLUS 算法基础上进行改进，提出 NNSB-OPTICS 聚类算法，该算法首先设计了一种全局的最近邻指针，该指针始终指向种子队列中最近邻距离最小的点，算法完成一次扩张后取出最近邻指针指向的点进行下一次迭代，不需要进行排序操作，有效提高时间效率，其次，提出一种点平均距离的概念，通过迭代扩张获取每一个数据对象的点平均距离，形成包含数据集聚类结构信息的点平均距离排序，根据点平均距离排序进行可将数据集划分为若干类簇，避免了用户设置阈值，降低了人为因素对滑坡预测结果的影响。为方便叙述，对于给定数据集 $X=\{x_1,x_2,\cdots x_n\}$ 首先给出如下定义：

定义 1　最近邻距离。设存在两个集合 M 和 N，$M\cup N=X$，$M\cap N=\varnothing,\forall m\in M$ 的最近邻距离为：

$$ND_m=\begin{cases}0, & ,N=\varnothing\\ \min\limits_{n\in N}\{DIS(m,n)\}, & N\neq\varnothing\end{cases}$$

定义 2　点平均距离。设 $\forall x_i\in X$，x_i 的点平均密度为：

$$ADP_i=\frac{\sum_{j=1}^{n}dist(x_i,x_j)}{n}$$

其中 $dist(x_i,x_j)$ 为 X 中 x_i 与 x_j 的距离。

NNSB-OPTICS 聚类算法通过最近距离迭代扩张，生成一个点平均距离排序队列，分析点平均距离排序队列中点平均距离的陡峭上升和下降区域，可以有效地区分数据点密集区和稀疏区，从而将数据集划分为若干类簇，避免了人为设置密度阈值来划分密集区和稀疏区。在跌待扩张过程中，为避免多次排序和重复计算相似度降低时间效率，NNSB-OPTICS 算法在 OPTICS-PLUS 算法基础上作出以下两点改进：a)定义了一个 GPNP(global point to nearest point)指针，如图 1 所示，每一次迭代都记录最近邻距离最小的点，并将 GPNP 域指向该点，下一次迭代开始时，直接取出 GPNP 域指向的点进行扩张；b)NNSB-OPTICS 算法为每个对象额外添加一个 SD(Sum of Distance)域，记录该对象与已扩张对象的总距离，求取某对象点平均距离时，直接读取 SD 域中的总距离。

NNSB-OPTICS 算法在 OPTICS-PLUS 算法基础上对 OPTICS 算法的扩张策略进行优化，在迭代扩张过程中并不会对已扩张的数据对象进行比较，每一次迭代都将减少一个比较对象，因此算法时间复杂度为 $T(n)=O(\sum_{i=1}^{n}i)=O(\frac{1}{2}n^2+\frac{1}{2}n)$，即

NNSB-OPTICS 算法时间复杂度为 $O(n^2)$，与 OPTICS 算法时间复杂度[13]相等。但在确定算法扩张方向时，OPTICS 算法需要对有序种子队列进行排序，该过程时间复杂度为 $O(r^2)$（r 为种子队列长度）[14]，NNSB-OPTICS 算法只需根据 GPNP 指针进行搜索，取出 GPNP 指针指向的数据对象进行扩张，时间复杂度为 $O(r)$，因此 NNSB-OPTICS 算法实际效率要高于 OPTICS 算法。

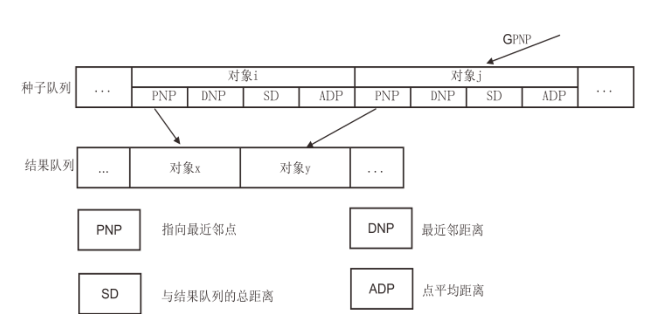


图 1 NNSB-OPTICS 算法数据对象存储结构

1.2 不确定数据处理

滑坡的发生受多种因素的影响，其中降雨是重要的诱发因素，但降雨量的取值属于不确定数据，只能确定其大概取值范围，无法精确描述其数值大小[14]。本文提出的 NNSB-OPTICS 聚类算法适用于连续性和离散型属性数据，但不确定数据降雨量仍不能进行有效刻画及处理。为此本文引入 EW 型距离公式和云模型理论，并根据降雨量数据的分布特征，将 EW 型距离公式与云模型理论相结合，得到一种新的不确定数据距离公式。

定义 3 不确定数据[15]。设存在映射 f 使得 $x\in V=[v^-,v^+]$, $v^+>v^-$ 有：$f(x)\in[0,1]$，且 $\int_{-\infty}^{v^-}f(x)=0$，$\int_{v^-}^{v^+}f(x)=1$，$\int_{v^+}^{+\infty}f(x)=0$ 则，称 x 为不确定数据，V_k 为 x 的取值区间，$f(x)$ 为 x 的概率密度函数。

在不确定数据的距离度量方面， EW 型距离是不确定数据距离度量中使用最为广泛的方法。对于给定不确定数据 $x_i\in[v_i^-,v_i^+]$ 和 $x_j\in[v_j^-,v_j^+]$，x_i 与 x_j 间的 EW 型距离为

$$dist_{EW}^p(x_i,x_j)=\sqrt[p]{\left|E(x_i)-E(x_j)\right|^p+\frac{1}{3}\left|W(x_i)-W(x_j)\right|^p},p\ge 1$$

其中：$E(x_i)=(v_i^-+v_i^+)/2$ 和 $E(x_j)=(v_j^-+v_j^+)/2$ 分别为 x_i 和 x_j 的期望值，$W(x_i)=(v_i^+-v_i^-)/2$ 和 $W(x_j)=(v_j^+-v_j^-)/2$ 分别为 x_i 和 x_j 的宽度。

EW 型距离认为不确定数据在取值区间内服从均匀分布，从而综合了不确定数据的期望和区间宽度来度量不确定数据间的距离。但本文研究的不确定数据降雨量在其取值区间内近似服从正态分布[16]，故 EW 型距离不能直接用于降雨量刻画中。根据延安气象调查局数据显示，相邻地域降雨具有相似的降雨量，根据这一性质，根据绥德气象调查局数据显示，相邻地域降雨具有相似的降雨量，根据这一特性，可以采用逆向云算法[17]，获取降雨量 x_i 对应的正态云模型数字特征，对 x_i 进行定性描述。抽取 x_i 邻近的 $l(l=1,2,3K)$ 个不确定数据取值区间作为正态模糊数的 α- 水平截集进行模糊化，得到模糊隶属函数曲线；根据扩张原理将模糊隶属度函数曲线扩张成云期望曲线，曲线方程为 $\mu=\exp-(\frac{x-a}{\sqrt{2}\sigma})^2$；从云期望曲线中得出云模型数字特征，其中期望 Ex 和超熵 He 计算公式如下：

$$Ex(x_i)=a \tag{1}$$

$$He(x_i)=\sqrt{\frac{\sum_{i=i-l/2}^{i+l/2}(\sigma_i'-\bar{\sigma}_i')^2}{l}} \tag{2}$$

其中：a 为云期望曲线方程的均值，σ_i' 为云期望曲线的方差，$\bar{\sigma}_i'$ 为 σ_i' 的均值。

将云模型期望和超熵引入到 EW 型距离中，以超熵代替 EW 型距离中的取值区间宽度对不确定数据的模糊性进行描述，得 EC 型距离，公式如下：

$$dist_{EC}^p(x_i,x_j)=\sqrt[p]{\left|Ex(x_i)-Ex(x_j)\right|^p+\frac{1}{3}\left|He(x_i)-He(x_j)\right|^p},p\ge 1$$

对于离散属性和连续属性，其取值不具有模糊性，直接令其期望等于属性值，超熵等于 0，然后使用 EC 型距离公式进行距离度量，故 EC 型距离公式适用于包含离散值属性、连续值属性和不确定属性的数据集。

1.3 不确定 NNSB-OPTICS 聚类算法设计

以 EC 型距离公式作为相似度计算公式，应用到 NNSB-OPTICS 聚类算法中，提出不确定 NNSB-OPTICS 聚类算法。不确定 NNSB-OPTICS 算法具体流程如下所示。

算法：NNSB-OPTICS 算法

输入：数据集 $X=\{x_1,x_2,\cdots,x_n\}$

输出：结果队列 $OrderList$

（1） $FOR\ ALL\ p\in X\ DO$

（1.1）计算 p 各属性的期望和超熵

（2） $OrderList\leftarrow\varnothing$，令 $GPNP$ 指向任意点

（3） $WHILE(X\ne\varnothing)$

（3.1）从 X 中取出 $p=x_{GPNP}$，并令 $X:=X\setminus\{x_{GPNP}\}$

（3.2） $IF(X=\varnothing)$

（3.2.1） $ADP_p\leftarrow SD_p/n$，将 p 添加到 $OrderList$ 末尾

（3.3） $FOR\ ALL\ q\in X\ DO$

（3.3.1）计算 p 与 q 的距离 $dist_{EC}(p,q)$，并令 $SD_p\leftarrow SD_p+dist_{EC}(p,q)$，$SD_q\leftarrow SD_q+dist_{EC}(p,q)$

（3.3.2） $IF(dist_{EC}(p,q)<DNP_q)$

（a） $DNP_q\leftarrow dist_{EC}(p,q)$

（b）令 PNP_q 指向 p

（c）令 $GPNP\leftarrow PNP_p$

（3.4）令 $ADP_p\leftarrow SD_p/n$，将 p 加入到 $OrderList$ 末尾

（4）根据 PNP 指向对 $OrderList$ 进行重组

（5） $RETURN\ OrderList$，算法结束

2 实验结果与分析

2.1 实验环境

本文实验均在 Win7 操作系统，Intel(R) Core(TM) i5-4210U 2.80GHz CPU，8G 内存计算机中进行。滑坡实验数据使用 ARCGIS10.3 软件提取，数据库平台为 Oracle 12c，算法由 Python 语言在 PyCharm5.03 平台中测试。

2.2 仿真实验

为了验证本文算法的聚类有效性，将 NNSB-OPTICS 算法分别与 OPTICS 算法[12]、OPTICS-PLUS 算法[9]、和 EOPTICS 算法[18]在 4 个 UCI 数据集上进行对比实验，数据集特征如表 1 所示。实验主要针对算法聚类准确率、聚类结果稳定性和时间效率等方面进行测试。

表 1 实验选用 UCI 数据集特征

数据集	样本数量	数据维数	类别数
Iris	150	4	3
Wine	178	13	3
Seed	210	7	3
Balance	625	4	3

本文采用 Micro-precision 标准，利用数据分类信息来衡量聚类结果准确率，计算公式如下：

$$MP = \frac{1}{N}\sum_{h=1}^{h=k} a_h$$

其中：a_h 表示正确聚类的样本点数量，N 为数据集样本点总数，k 表示聚类数量。MP 的值在区间[0,1]内，MP 的值越接近于 1 聚类准确率越高。实验前首先通过多次测试获取 OPTICS-PLUS 算法在数据集中聚类结果最优时的领域半径 $MinPts$ 和核心点数 Eps，然后设定实验中领域半径的取值集合为 $\{MinPts-2, MinPts, MinPts+2\}$，领域半径取值集合为 $\{Eps-0.15, Eps, Eps+0.15\}$，最后对领域半径和核心点数进行组合，获得 9 组参数，每组参数运行 10 次，共计 90 次实验。为了更好的分析算法的准确率和稳定性，记录最优结果 MP_{max} 和最差结果 MP_{min}，并计算多次实验的均值 MP_{mean}，MP_{mean} 计算公式如下：

$$MP_{mean} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{h=1}^{h=k} a_h$$

其中：T 为实验重复次数，本文实验中 T =90。各算法在 UCI 数据集上聚类准确率和单次运行时间如表 2、3 所示。

由表 2 的实验对比结果可以看出，在 4 个 UCI 数据集上，NNSB-OPTICS 聚类算法的平均准确率均要优于其他三个算法，从最优结果和最差结果差值的对比上看，NNSB-OPTICS 聚类算法差值明显较小，说明 NNSB-OPTICS 聚类算法聚类准确率较高且聚类结果稳定性较好，原因有二：其一，NNSB-OPTICS 算法避免了人为设置密度阈值，减小了人为因素对聚类结果的影响；其二，NNSB-OPTICS 算法扩张完成后，根据最近邻指向进行结果重组织策略将边界点划分到最近邻的密集区中，提高了聚类准确率。但是从最优结果显示，在 Iris 数据集和 Wine 数据集上，不确定 NNSB-OPTICS 算法的最优结果要低于 OPTICS-PLUS 聚类算法，这是因为 Iris 数据集的 1 类样本点与 2 类样本点、Wine 数据集 3 类样本点均存在数据交叉，且交叉部分数据密度分部变化较小，导致 NNSB-OPTICS 聚类算法生产的点平均距离排序波动平缓，类簇识别效果较差，这是 NNSB-OPTICS 聚类算法需要进一步改进的地方。从表 3 中各算法时间效率比较上看，NNSB-OPTICS 聚类算法单次运行时间时间最少，时间效率上具有明显的优势。

表 2 各算法聚类准确率比较/%

算法	Oris			Wine		
	MPmax	MPmin	MPmean	MPmax	MPmin	MPmean
OPTICS	77.52	69.85	74.82	85.5	80.33	83.46
EOPTICS	77.86	72.39	75.72	87.13	83.08	85.08
OPTICS-PLUS	80.36	75.61	77.23	88.49	83.21	85.75
NNSB-OPTICS	78.64	76.32	77.9	88.21	86.45	87.72
算法	Balance			Seed		
	MPmax	MPmin	MPmean	MPmax	MPmin	MPmean
OPTICS	82.43	75.38	78.55	83.01	76.33	79.92
EOPTICS	83.59	76.27	80.16	85.72	77.6	81.54
OPTICS-PLUS	84.4	79.68	82.71	86.97	82.17	85.16
NNSB-OPTICS	84.67	81.31	83.74	87.36	84.93	86.28

表 3 各算法单次运行时间比较/ms

算法	Iris	Wine	Seed	Balance
OPTICS	386	473	672	4430
OPTICS-PLUS	437	551	804	5618
EOPTICS	405	527	766	5114
NNSB-OPTICS	45	138	175	436

在 UCI 数据集上的实验结果表明，相比 OPTICS 聚类算法、OPTICS-PLUS 聚类算法和 EOPTICS 聚类算法，NNSB-OPTICS 聚类算法能够更加有效地避免人为因素对聚类结果的影响，聚类稳定性较高，且时间效率较优。

2.3 实例应用

为了验证不确定 NNSB-OPTICS 算法在滑坡危险性预测中是否具有可行性以及本文提出的不确定数据处理方法能否有效刻画降雨量，选取延安市宝塔区作为研究区进行实例验证。延安市宝塔区地处陕北黄土高原中部，地质条件复杂，人类活动频繁，降雨等影响因素导致滑坡发生频率较大，人类生命安全和财产受到巨大的威胁。降雨不确定因素在实际滑坡危险性预测过程中难以有效刻画，根据本文不确定数据处理的方式对降雨进行处理，结合滑坡相关理论基础，将不确定 NNSB-OPTICS 聚类算法应用到宝塔区滑坡危险性预测研究中，验证不确定 NNSB-OPTICS 聚类算法在滑坡危险性预测中的可行性。

2.3.1 数据来源及数据预处理

本文以宝塔区地质灾害详细调查项目为背景，进行滑坡危险性预测研究。实验数据来源如下：采用 ARCGIS 软件将延安市宝塔区用栅格划分模块，划分网格尺寸大小为 5m×5m，把研究区划分为 5672922 个网格单元。每个网格单元看成一个点，导入到精度为 1:5000 数字高程图中，派生出坡型、坡度、坡高和坡向专题图，再从这些专题图中分别获取坡型、坡度、坡高和坡向的数据信息；岩土体结构数据从 1:10000 地质图中获取；植被覆盖数值来源于全区 Spot5 遥感数据，采用 ERDAS 遥感影像处理软件从 Spot5 近红外波段 B3 和可见光红波段 B2 进行归一化差值计算获取；降雨量值是采用滑坡发生前 7d 的 24h 降雨量以及从气象雨量图中获取未来 7d 的 24h 降雨量。

获取的原始数据集中样本数目多达数百万条，包含属性项众多，其中包含大量缺失值、重复值、错误值。为了提高实验结果的准确率，需要对原始数据集进行数据预处理操作。首先结合滑坡相关理论以及黄土高原特殊的地质环境发生灾害的特征进行数据降维，保留坡度、坡高、坡向、植被、坡型、岩土体结构、降雨等 7 个属性项作为聚类特征属性，滑坡危险等级作为决策属性，删除其余对滑坡发育影响较小的属性项。然后进行数据清洗，删除包含缺失值、重复值、错误值的记录。经过数据预处理后，获得有效的记录数据 5647382 条，数据集中属性特征如表 4 所示。

表 4　滑坡数据集属性特征

属性项	属性类型	离散型属性取值
坡度	连续	
坡高	连续	
坡向	连续	
植被	离散	低，较低，高，较高
坡型	离散	凹型，凸型，阶梯型，直线型
岩土体结构	离散	黄土+近于水平古土壤层型， 黄土+倾斜古土壤层型 黄土+古土壤+基岩型， 黄土+古土壤+新近纪泥岩型
降雨量	不确定	
滑坡危险等级	离散	低危，中危，高危

2.3.2 不确定 NNSB-OPTICS 聚类算法模型构建

首先按照式(1)(2)计算数据集中各对象的云模型均值和超熵，然后初始化结果队列为空，从数据集中任意对象出发，通过本文提出的 EC 型距离度量公式计算新各对象之间的距离，一次迭代完成后按照 GPNP 指针的指向进行下一次迭代扩张，直到数据集中所有对象加入到结果队列中，获得一张包含点平均距离排序的点平均距离排序队列，最后采用 Gmdlent Clusterin 方法[19]，通过识别点平均距离排序中的陡峭上升和下降的区域进行类簇识别，输出聚类结果，获得 K 个类簇。

经过聚类后，滑坡样本数据集中所有评价单元被划分到 K 个类簇中，根据聚类算法性质可知，同一类簇内的对象具有较高的相似性，不同类簇间的对象具有较高的相异度，即同一类簇中的评价单元具有相似的地形地貌、气候环境特征。文献[20]证明，相似的滑坡发育特征具有相似的滑坡发生趋势，根据这一理论，利用野外勘测到的区域含有降雨信息的 293 个滑坡观测点的已知危险性等级，结合直接搜索法和专家评价法[21]可快速确定各类簇的滑坡危险性等级。通过直接搜索法，逐一搜索各类簇中的评价单元，若含有一个已确定的危险性等级时，类簇的危险性等级等同于类簇内评价单元的危险性等级，若含有两个及以上已确定的危险性等级且类簇内评价单元的危险性不同等级的数目不同时，类簇的危险性等级通过少数服从多数原则确定，若含有两个及以上已确定的危险性等级且类簇内评价单元的危险性不同等级的数目相同时或含有零个已确定的危险性等级，则通过滑坡灾害专家利用先前的滑坡预判经验以及对区域地质环境条件的熟知程度，结合区域地质调查结果判定滑坡危险性等级，从而划分出研究区剩余评价单元的危险性等级。

2.3.3 评价标准

通过对研究区滑坡实际调查数据和滑坡预测结果统计建立误差矩阵。在误差矩阵中，列表示实际观测值，矩阵行表示通过预测模型获得的预测值，例如预测值为低危，观测值是低危的样本数量用 P_{11} 表示，预测值为低危，观测值为中危的样本数量用 P_{21} 表示，预测值为低危，观测值为高危的样本数量用 P_{31} 表示。Kappa 系数[22]是一种较为简单、准确度较高的评价方法，基于误差矩阵的 Kappa 系数精度评价方法能够在统计意义上反映分类结果的优越性。Kappa 系数计算公式为

$$Kappa=\frac{Pr_0-\frac{\sum_{i=1}^{n}(P_{i+}\cdot P_{+i})}{N^2}}{1-\frac{\sum_{i=1}^{n}(P_{i+}\cdot P_{+i})}{N^2}} \tag{4}$$

$$Pr_0=\frac{\sum_{i=1}^{n}p_{ii}}{N}\cdot 100\% \tag{5}$$

其中：Pr_0 是预测模型的总体精度(overall accuracy)，表示数据集中预测值和观测值相一致的概率，P_{i+} 表示第 i 行记录总数，P_{+i} 表示第 i 列记录总数，N 为样本数量，n 为分类的类型数量，在本实验中 n 取值为 3。Kappa 系数取值在区间[0,1]中，当数据集中所有样本预测值与观测值完全吻合时，Kappa 系数的值为 1。

2.3.4 滑坡预测精度评价分析与比较

为验证本文提出的不确定数据处理方法是否能够有效处理降雨量数据，提高滑坡预测精度，分别使用 NNSB-OPTICS 聚类算法和不确定 NNSB-OPTICS 聚类算法建立滑坡预测模型，对延安市宝塔区进行滑坡危险性预测。对于不确定数据降雨量，不确定 NNSB-OPTICS 滑坡预测模型采用式（1）（2）获得降雨量云模型数字特征，然后采用式（3）进行相似度计算；NNSB-

OPTICS 聚类算法采用传统滑坡预测中处理降雨量所使用的定量法[23]进行离散化，即将降雨划分为以下几类：小雨（20 mm 以下），中雨（20~44.9 mm），大雨（45~59.9 mm），暴雨（60~79.9 mm），大暴雨（80~99.9 mm），特大暴雨（100 mm 以上）,并分别以数值代替，采用欧氏距离进行相似度计算。宝塔区地质灾害灾害观测点有 428 处，其中有滑坡观测点有 293 个，在数据预处理阶段，所有滑坡观测点被栅格化为 1367 个评价单元，其中包含低危评价单元 311 个，中危评价单元 729 个，高危评价单元 327 个。对 NNSB-OPTICS 滑坡危险性预测模型和不确定 NNSB-OPTICS 滑坡危险性预测模型预测结果进行统计，获得误差矩阵，如表 5 所示。

表 5　两种预测模型滑坡危险性预测误差矩阵

	预测观测	低危	中危	高危	预测总和
NNSB-OPTICS 聚类模型	低危	254	47	22	322
	中危	41	652	43	736
	高危	16	31	262	309
	观测总和	311	729	327	1367
	预测观测	低危	中危	高危	预测总和
不确定 NNSB-OPTICS 聚类模型	低危	263	31	12	306
	中危	35	674	26	735
	高危	13	24	289	326
	观测总和	311	729	327	1367

结合表 6 和式（4）（5）进行计算，可得 NNSB-OPTICS 算法聚类算法滑坡预测模型总体精度和 Kappa 系数为

$$\Pr_0 = \frac{254+652+262}{1367} \times 100\% = 85.4\%$$

$$Kappa = \frac{0.854-(322\times311+736\times729+309\times327)/1367^2}{1-(322\times311+736\times729+309\times327)/1367^2} = 0.76$$

不确定 NNSB-OPTICS 算法总体精度和 Kappa 系数为

$$\Pr_0 = \frac{263+674+289}{1367} \times 100\% = 89.7\%$$

$$Kappa = \frac{0.897-(306\times311+735\times729+326\times327)/1367^2}{1-(306\times311+735\times729+326\times327)/1367^2} = 0.83$$

计算结果表明，两种滑坡预测模型的总体精度都高于 80%，满足可信要求[24]，说明本文提出的 NNSB-OPTICS 聚类算法和不确定 NNSB-OPTICS 聚类算法在滑坡预测中是可行的。对比 NNSB-OPTICS 滑坡危险性预测模型和不确定 NNSB-OPTICS 滑坡危险性预测模型的总体精度和 Kappa 系数可知，不确定 NNSB-OPTICS 滑坡危险性预测模型的总体精度高出 NNSB-OPTICS 滑坡危险性预测模型 4.3 个百分点，Kappa 系数也大于 NNSB-OPTICS 滑坡危险性预测模型，证明在相同滑坡样本数据集下，不确定 NNSB-OPTICS 滑坡预测模型的预测结果更加吻合于滑坡实际情况，这是因为不确定 NNSB-OPTICS 滑坡预测模型在降雨量刻画上采用了 EC 型距离公式，充分考虑的降雨量分布特征，弥补了传统滑坡预测方法将降雨量直接离散化造成信息丢失的不足，能够更加有效刻画降雨量数据，在一定程度上提高了滑坡危险性预测精度。

3　结束语

传统聚类算法在滑坡危险性预测中不能有效刻画不确定数据降雨量且对分布不均匀的数据聚类效果较差，对此，本文首先引入一种基于密度的 OPTICS-PLUS 算法，并针对该算法需要人工设置密度阈值、时间复杂度高等不足，提出 NNSB-OPTICS 聚类算法，然后考虑的降雨量数据的分布特征，结合 EW 型距离公式和云模型理论提出 EC 型距离公式，将 EC 型距离公式应用到 NNSB-OPTICS 中，提出不确定 NNSB-OPTICS 聚类算法，解决了降雨量数据难以有效刻画及处理的问题。通过实例滑坡危险性预测证明本文提出的不确定数据处理方法能够更加有效刻画降雨量，提高滑坡预测精度。

参考文献：

[1] 黄润秋. 20 世纪以来中国的大型滑坡及其发生机制 [J]. 岩石力学与工程学报, 2007, 26 (3): 434-454

[2] 周涛, 陆惠玲. 数据挖掘中聚类算法研究进展 [J]. 计算机工程与应用, 2012, 48 (12): 100-111.

[3] Gorsevski P V, Gessler P E, Jankowski P. A fuzzy K-means classification and a bayesian approach for spatial prediction of landslide hazard [M]// Handbook of Applied Spatial Analysis. Berlin: Springer, 2010: 653-684.

[4] 胡凯衡, 崔鹏, 韩用顺, 等. 基于聚类和最大似然法的汶川灾区泥石流滑坡易发性评价 [J]. 中国水土保持科学, 2012, 10 (1): 12-18.

[5] 桂蕾, 殷坤龙, 王佳佳. 基于聚类分析的滑坡灾害危险性区划研究 [J]. 水文地质工程地质, 2013, 40 (1): 100-105.

[6] 张俊, 殷坤龙, 王佳佳, 等. 三峡库区万州区滑坡灾害易发性评价研究 [J]. 岩石力学与工程学报, 2016, 35 (2): 284-296

[7] Wang X, Niu R. Landslide prediction in three gorges based on cloud model and data field [C]// Proc of International Symposium on Computer Network and Multimedia Technology. 2009: 1-4.

[8] Wan S. Entropy-based particle swarm optimization with clustering analysis on landslide susceptibility mapping [J]. Environmental Earth Sciences, 2013, 68 (5): 1349-1366.

[9] 曾依灵, 许洪波, 白硕. 改进的 OPTICS 算法及其在文本聚类中的应用 [J]. 中文信息学报, 2008, 22 (1): 51-55.

[10] 包玉娥, 彭晓芹, 赵博. 基于期望值与宽度的区间数距离及其完备性 [J]. 模糊系统与数学, 2013, 27 (6): 133-139.

[11] 李德毅, 刘常昱. 论正态云模型的普适性 [J]. 中国工程科学, 2004, 6 (8): 28-34.

[12] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: ordering points to identify the clustering structure [J]. ACM SIGMOD Record, 1999, 28 (2): 49-60.

[13] Zhang Q, Wang X, Wang X. An OPTICS clustering-based anomalous data filtering algorithm for condition monitoring of power equipment [M]// Data

Analytics for Renewable Energy Integration. [S. l.] : Springer International Publishing, 2015: 917-922.

[14] 侯荣涛，路郁，王琴，等. OPTICS 算法在雷电临近预报中的应用 [J]. 计算机应用, 2014, 34 (1): 297-301.

[15] 毛伊敏，彭喆，陈志刚，等. 基于不确定决策树分类算法在滑坡危险性预测的应用 [J]. 计算机应用研究, 2014, 31 (12): 3646-3650.

[16] 刘卫明，高晓东，毛伊敏，等. 不确定遗传神经网络在滑坡危险性预测中的研究与应用 [J]. 计算机工程, 2017, 43 (2): 308-316.

[17] 于少伟，史忠科. 基于正态分布区间数的逆向云新算法 [J]. 系统工程理论与实践, 2011, 31 (10): 2021-2026.

[18] Alzaalan M E, Aldahdooh R T, Ashour W. EOPTICS: enhancement ordering points to identify the clustering structure [J]. International Journal of Computer Applications, 2012, 40 (17): 975-8887.

[19] Brecheisen S, Kriegel H P, Kroger P, et al. Density-based data analysis and similarity search [M]// Multimedia Data Mining and Knowledge Discovery. 2007: 94-115.

[20] Kwang Y, Jong H. Landslide susceptibility mapping in Injae, Korea, using a decision tree [J]. Engineering Geology, 2010, 116 (3): 274-283

[21] Guzzetti F, Carrara A, Cardinali M, Reichenbach P. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy [J]. Geomo-rphology, 1999, 31 (1-4): 181-216

[22] 许文宁，王鹏新，韩萍，等. Kappa 系数在干旱预测模型精度评价中的应用——以关中平原的干旱预测为例 [J]. 自然灾害学报, 2011. 20 (6): 81-86.

[23] 高华喜，殷坤龙. 降雨与滑坡灾害相关性分析及预警预报阈值之探讨 [J]. 岩土力学, 2007, 28 (5): 1056-1060

[24] Sabokbar H F, Roodposhti M S, Tazik E. Landslide susceptibility mapping using geographically-weighted principal component analysis [J]. Geomorphology, 2014, 226 (1): 15-24.